Previsão do Coeficiente de Permeabilidade Através de Inteligência Artificial

Lucas Marques Pires Da Silva
Engº Geotécnico, COPPE-UFRJ, Rio de Janeiro, Brasil, lucas.silva@coc.ufrj.br

Gleyce de Souza Baptista
Engª Geotécnica, PUC-Rio, Rio de Janeiro, Brasil, gleycesouzaa@gmail.com

Gustavo Vaz de Mello Guimarães, DSc
Professor, UFRJ/Macaé, Macaé/RJ, Brasil, gvmg@poli.ufrj.br

RESUMO: O presente estudo busca desenvolver um software com base em inteligência artificial para a estimativa do coeficiente de permeabilidade através de granulometria e índices físicos do solo/rochas. O banco de dados utilizado foi compilado por Feng *et. al*. (2023) e disponibilizado pelo ISSMGE. Um total de 59 tipos diferentes amostras em mais de 50 pesquisas distintas estão sendo analisadas. Destaca-se que 703 amostras possuem granulometria completa e todos os índices físicos. Por outro lado, 575 amostras, que também estão sendo utilizadas, não possuem todas as informações sobre granulometria e/ou índices físicos. Os algoritmos estão sendo desenvolvidos utilizando a granulometria, coeficiente de uniformidade e curvatura, densidade real dos grãos e índice de vazios para prever o coeficiente de permeabilidade de quase 1300 amostras de solo/rochas. Os resultados parciais mostram que uma IA já desenvolvida fornece um bom desempenho para amostras com permeabilidades elevadas. Porém os resultados desta IA ainda não são satisfatórios para amostras com permeabilidades menores. Em continuidade a pesquisa, espera-se desenvolver outra IA para solos com permeabilidades menores, tendo então um único software com duas inteligências artificias que funcionarão para diferentes espectros de granulometrias.

PALAVRAS-CHAVE: Coeficiente de Permeabilidade, Inteligência Artificial, Granulometria, Índices Físicos

ABSTRACT: The current study aims to develop artificial intelligence-based software for estimating the permeability coefficient using soil/rock grain size distribution and physical indices. The database employed was compiled by Feng *et al*. (2023) and made available through ISSMGE. A total of 59 different sample types from over 50 distinct studies are being analyzed. It is noteworthy that 703 samples have complete grain size distribution and all physical indices. Conversely, 575 samples, which are also being used, lack complete information on grain size distribution and/or physical indices. The algorithms are being developed using grain size distribution, coefficient of uniformity and curvature, actual grain density, and void ratio to predict the permeability coefficient of nearly 1300 soil/rock samples. Preliminary results indicate that an already developed AI exhibits good performance for samples with high permeabilities. However, the results of this AI are not yet satisfactory for samples with lower permeabilities. As research progresses, another AI is expected to be developed for soils with lower permeabilities, resulting in a single software incorporating two artificial intelligences, each functioning for different grain size distribution spectrums.

KEYWORDS: Permeability Coefficient, Artificial Intelligence, Grain Size Distribution, Physical Indices

1 INTRODUÇÃO

O estudo da permeabilidade é essencial para engenheiros civis, segundo Harr (1991) os problemas que envolvem o fluxo de água subterrâneas envolvem: estimativa da quantidade de fluído percolado, definição do domínio de fluxo e análise de estabilidade. Cedergren (1989)menciona que a maioria dos engenheiros experientes em percolação considera a teoria da percolação como um meio de prever a ordem de grandeza dos problemas e de desenvolver tipos apropriados de soluções. Um fator importante para a análise de fluxo é o coeficiente de permeabilidade.

O coeficiente de permeabilidade é um parâmetro que integra parâmetros do fluído e do material pelo qual o fluído permeia. A eq. (1), Mitchell e Soga (1930), mostra uma formulação para estimar o coeficiente de permeabilidade sob a hipótese de grãos perfeitamente esféricos.

$$k = \frac{\gamma_w}{\mu} \cdot C \cdot D_S \cdot \frac{e^3}{1+e} \cdot S^3 \quad (1)$$

Note que a equação 1 mostra a influência de diversos parâmetros da permeabilidade do solo de forma direta como o peso específico (γ_w) e a viscosidade (μ) do líquido, o coeficiente de forma (C), diâmetro representativo (D_S) e índice de vazios (e) do material granular e a saturação (S) do sistema.

Existem outras formulações como a de Hazen (1892) que correlaciona a permeabilidade com a o D_{10} do solo, de acordo com essas formulações percebe-se que a saturação, índice de vazios, o granulometria e o formato dos grãos são fatores que mais influenciam na permeabilidade intrínseca do solo. Hazen (1892) indicou que a permeabilidade é proporcional ao quadrado do tamanho efetivo de grão para a areia fofas com partículas uniformes.

A Figura 1 apresenta a permeabilidade como uma estimativa da ordem de grandeza, relacionada com a granulometria do material (Mello e Teixeira, 1962).

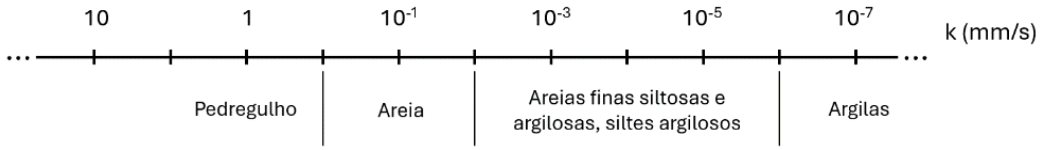

Figura 1 Classificação granulométrica segundo o coeficiente de permeabilidade (Adaptado de Mello e Teixeira, 1962)

Rawls e Brakensiek (1989) apresentam uma metodologia para a previsão dos coeficientes de permeabilidade do solo, estimando parâmetros para as funções de retenção de água e permeabilidade no solo. Os autores propuseram uma regressão considerando a porosidade, percentual de argila e areia em solos não saturados.

Sperry e Peirce (1995) desenvolveram um método para de quantificar as formas das partículas e determinar como essas formas, juntamente com o tamanho e a porosidade, influenciam a condutividade hidráulica através de experimentos com esferas de vidro, areia e partículas de vidro de formas irregulares.

No estudo de Lebron *et al*., (1999), utilizaram uma função de correlação e *software* de análise de imagem para examinar microfotografias para analisar o espaço poroso e geometria dos poros do solo. Além disso, o estudo buscou interpretar esses dados empregando redes neurais para a predição das propriedades hidráulicas do solo.

Conforme contextualizado por diversos autores, como por exemplo Mitchell e Soga (1930) e Hazen (1892) citados anteriormente, é possível perceber que existem diversos fatores que podem influenciar a permeabilidade de forma direta, conforme já abordado, ou indireta, como a temperatura, composição do solo, arranjo dos grãos, dentre outros. O estudo indicou que estas relações empíricas têm certas limitações, bem como incertezas. A IA surge, nesse contexto, como uma forma de ampliação das capacidades analíticas, além das técnicas tradicionais baseadas em propriedades do solo mais generalizadas. De acordo com Kanungo *et al*. (2014), técnicas de Inteligência Artificial, têm sido mais utilizadas recentemente para resolver problemas variados em geociências e engenharia geotécnica.

Com efeito, a fim de contribuir com a discursão do tema, foi desenvolvido um modelo preditivo baseado em técnicas de aprendizado de máquina, como KNN *(K-Nearest Neighbors*), *RF (Random Forest)* e ANN *(Artificial Neural Network)*, para estimar o coeficiente de permeabilidade do solo.

2 MATERIAIS E MÉTODOS

2.1 Materiais

O banco de dados analisado contém 59 tipos de solos com diferentes frequências de ocorrência. A mistura de areia e silte 50:50 é a mais comum, contabilizando 108 ocorrências. Seguida pela categoria de *Offshore Sands-Shallow Marine Carbonates* registrando 90 ocorrências. Além desses, existem diversas outras categorias com diferentes composições, a Tabela 1 apresenta alguns dados estatísticos do banco de dados. Os dados foram compilados por Feng *et al*. (2023) e estão disponíveis através da *International Society of Soil Mechanics and Geotechnical Engineering* (ISSMGE).

Tabela 1 Panorama dos dados presentes no banco de dados

Parâmetro	mín	média	máx
Gs	2,320	2,670	3,710
D10 (mm)	0,00	0,644	11,000
D20 (mm)	0,015	1,000	12,700
D25 (mm)	0,017	1,200	13,100
D30 (mm)	0,022	1,420	13,600
D40 (mm)	0,045	1,900	19,400
D50 (mm)	0,077	2,360	24,100
D60 (mm)	0,081	2,910	27,500
D70 (mm)	0,085	3,610	30,100
D75 (mm)	0,091	4,020	37,400
D90 (mm)	0,143	5,850	67,600
CU	1,200	16,600	$1{,}27 \times 10^{3}$
CZ	0,094	3,680	615,001
e	0,111	0,568	1,470
k (mm/s)	$9{,}78 \times 10^{7}$	7,020	561,010
Log_k (mm/s)	-6,01	-1,320	2,750

2.2 Métodos

A análise de dados foi realizada utilizando a linguagem de programação *Python*, utilizando a biblioteca *scikit-learn.* Inicialmente, o conjunto de dados foi pré-processado para garantir a qualidade e a relevância das informações para o treinamento dos modelos. As variáveis independentes utilizadas para a previsão (X) e a variável dependente (y) foram separadas, removendo-se do conjunto de preditores a variável dependente e outras que não são essenciais para o modelo.

O conjunto de dados foi dividido de forma aleatória em partes de treinamento e teste, com 80% dos dados alocados para treinamento e 20% para teste. Para mitigar a influência de diferentes escalas entre as variáveis, aplicou-se a normalização dos dados utilizando o método *StandardScaler* do *scikit-learn.* Três modelos de aprendizado de máquina foram implementados e avaliados.

2.2.1 K-Nearest Neighbors

O algoritmo do K-Vizinhos Mais Próximos (KNN, do inglês "*K-Nearest Neighbors*") é explicado por James et. al., (2023) como um método de classificação que se baseia na proximidade entre as amostras. Selecionando um número K de vizinhos mais próximos, o algoritmo identifica os K dados de treino mais semelhantes à nova observação. A classificação é então feita com base na classe predominante entre esses vizinhos. Simplificadamente, se uma nova observação é cercada por vizinhos de uma determinada classe, o KNN atribui essa classe à observação. A fronteira de decisão do KNN delimita as regiões onde as observações são classificadas em uma classe ou outra, baseando-se nesse princípio de maioria.

O método é exemplificado na Figura 2 em um cenário simples com doze observações, divididas igualmente entre duas classes, representadas por cores azul e laranja. A técnica é ilustrada considerando K=3.

Um ponto de teste, para o qual se deseja prever a etiqueta de classe, é representado por uma cruz preta. Os três pontos mais próximos ao ponto de teste são identificados; a previsão é de que o ponto de teste pertença à classe mais frequente entre seus K vizinhos mais próximos, que neste caso é a classe azul. O contorno decisório do KNN, representado pela linha preta no exemplo, demarca as regiões de classificação: a grade azul denota a área na qual uma observação de teste seria atribuída à classe azul, e a grade laranja, à classe laranja. Este modelo foi configurado com parâmetros padrão, utilizando a distância ponderada inversamente pelo inverso da distância e o algoritmo de árvore de esfera para computar os vizinhos mais próximos.

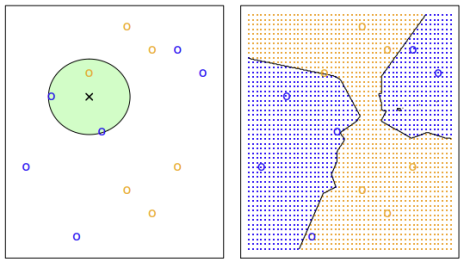

Figura 2 Exemplo simples de KNN utilizando K=3 (James *et al.*, 2023)

2.2.2 Random Forest

De acordo com Géron (2019), as Florestas Aleatórias (RF, do inglês *Random Forest*) funcionam com o treinamento de muitas Árvores de Decisão, exemplo de uma árvore na Figura 3, em subconjuntos aleatórios das características, e em seguida calculam a média de suas previsões. O autor pontua ainda que o RF está entre os algoritmos mais poderosos de Aprendizado de Máquina disponíveis atualmente. O RF geralmente é treinado pelo método *bagging* (*bootstrap aggregating*) com tamanho da floresta ajustada ao tamanho do conjunto de treinamento. O modelo de RF foi construído, utilizando 100 árvores de decisão e a capacidade de configurar o número mínimo de amostras por folha, entre outros parâmetros.

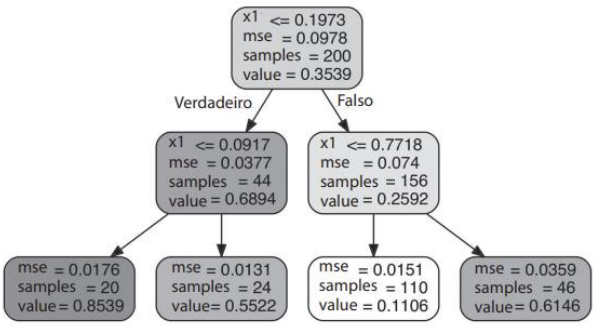

Figura 3 Exemplo de uma árvore de decisão de regressão (Géron, 2019)

2.2.3 Artificial Neural Network

Uma Rede Neural Artificial (ANN, do inglês "*Artificial Neural Network*") é um sistema que imita, de forma simplificada o funcionamento do cérebro para processar informações. A Figura 4 mostra um neurônio artificial básico, que recebe várias entradas, multiplica cada uma por um peso e soma esses valores. Depois, aplica uma função simples que determina se o neurônio se "ativa" ou não. Esse processo básico permite que a RNA aprenda a realizar tarefas complexas ao ajustar os pesos com base nos dados (Géron, 2019). Uma rede neural foi desenvolvida com uma camada oculta contendo 50 neurônios, função de ativação tangente hiperbólica, otimizador ADAM e uma taxa de aprendizado constante.

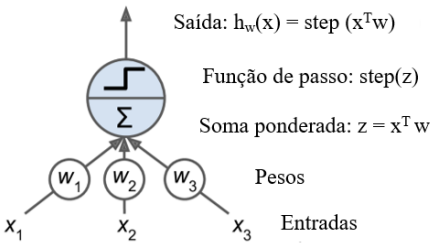

Figura 4 Exemplo simples de uma ANN (Géron, 2019)

Cada modelo foi treinado usando o conjunto de dados normalizados. A função *fit* foi utilizada para treinar os modelos nos dados, de modo que pudessem aprender a relação entre as variáveis independentes e a variável dependente.

A escolha dos modelos e seus respectivos parâmetros foi baseada em práticas comuns na literatura, visando uma comparação balanceada entre métodos tradicionais e algoritmos mais complexos de aprendizado de máquina. Os modelos KNN e SVR são exemplos de técnicas mais simples e compreensíveis, enquanto a ANN representa métodos mais sofisticados. O processo de treinamento foi realizado individualmente para cada modelo.

Após o treinamento, foi realizada a avaliação de desempenho dos modelos. As métricas utilizadas para mensurar a precisão das previsões em relação aos dados observados foram: o Erro Médio Absoluto (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o coeficiente de determinação (R^2). A primeira, identificada na eq. (2), é responsável por quantificar a média dos desvios absolutos entre os valores previstos e os observados. Por outro lado, a segunda métrica, expressa na eq. (3), é derivada pela obtenção da raiz quadrada da média dos quadrados dos erros, conferindo uma ponderação adicional aos desvios mais significativos. Essas ferramentas analíticas são fundamentais para avaliar a concordância entre o conjunto de projeções gerado pelo modelo e os dados reais, conforme descrito por Géron (2019). O coeficiente de determinação (R^2), representado na eq. (4), é uma medida estatística que representa a relação linear entre o valor real e o valor previsto, apresentando-se sempre entre 0 e 1 e sendo independente da escala da variável dependente (James *et al*., 2023). Esta medida estatística também foi utilizada para avaliação do desempenho do modelo entre os valores previstos pelo modelo e os dados reais.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - h(x_i)| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{h(x_i)}\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (4)$$

3 RESULTADOS

3.1 Análise do banco de dados

Na fase de análise do banco de dados, decidiu-se pela utilização do logaritmo do coeficiente de permeabilidade (k) para fins de treinamento e teste dos modelos. Esse tratamento é justificado pela natureza dos dados, que pode ser observado na Figura 5.

Inicialmente, o conjunto de dados continha 1275 registros. Durante o pré-processamento valores atípicos (*outliers*) foram excluídos, baseando-se na escala de permeabilidade definida por Mello e Teixeira (1962), conforme ilustrado na Figura 5. Adicionalmente, registros incompletos ou com valores ausentes também foram removidos. Após essa etapa de tratamento, o banco de dados foi consolidado com um total de 703 entradas válidas antes da separação dos dados em grupos de treino e teste, que serão empregadas para desenvolvimento e validação dos algoritmos preditivos.

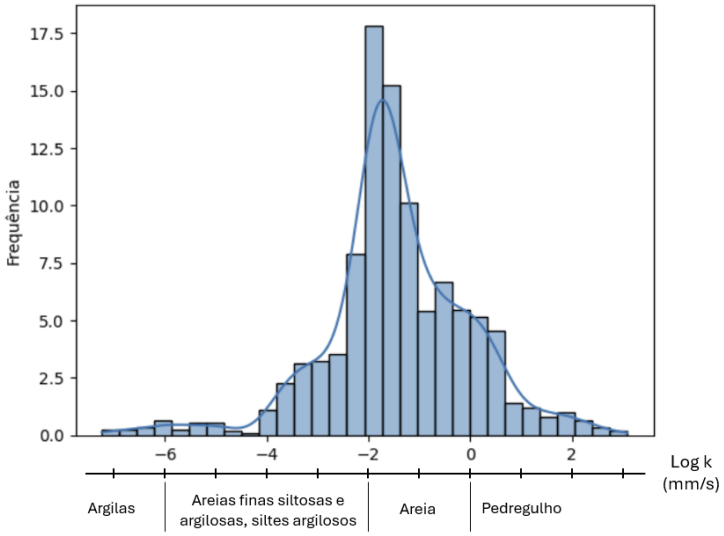

Figura 5 Histograma de distribuição dos valores de Log k (mm/s)

Aprofundando a investigação da associação entre as variáveis, procedeu-se com uma análise empregando o Coeficiente de Correlação de Pearson (r). Este coeficiente quantifica a intensidade e a direção da relação linear entre duas variáveis, variando entre -1 e 1. Conforme Filho e Júnior (2009), coeficientes próximos de 1 ou -1 indicam uma forte correlação linear, enquanto coeficientes próximos de zero sugerem uma correlação linear fraca entre as variáveis em análise.

A Figura 6 apresenta a matriz de correlação de Pearson para os dados estudados, detalhando as inter-relações entre diversos parâmetros. As barras azuis indicam a presença de correlações positivas, com coeficientes variando de 0,1 a 0,5, enquanto uma barra vermelha destaca uma correlação negativa modesta de -0,2. Observa-se que a maioria das variáveis apresenta uma correlação positiva de magnitude fraca a moderada, com a exceção de duas variáveis que mostram uma correlação negativa, indicando uma tendência inversa.

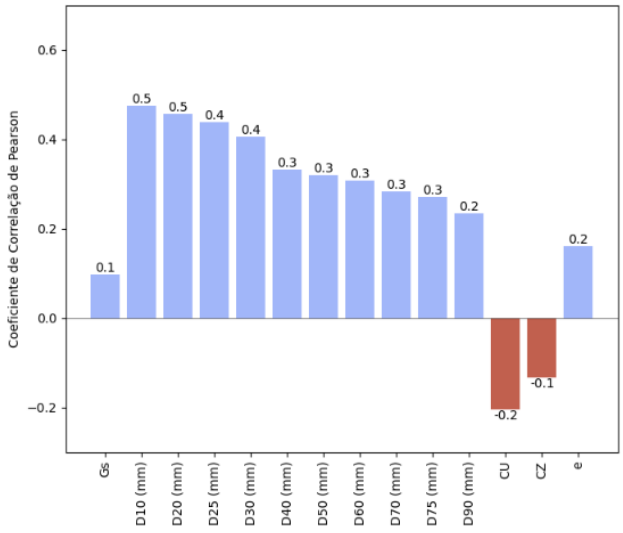

Figura 6 Matriz de correlação entre os parâmetros

3.2 Treinamento e avaliação dos algoritmos

A avaliação de desempenho dos modelos baseada nas métricas de avaliação (MAE, RMSE e R^2) é apresentada na Tabela 2. Observando os resultados de treino e teste, o modelo KNN apresentou um desempenho perfeito durante o treinamento com valores de MAE e RMSE iguais a zero e um R^2 de 1,00, o que pode sugerir uma tendência ao *overfitting* durante o treinamento, visto que houve uma diminuição

significativa do desempenho no conjunto de teste, com MAE e RMSE aumentando e R^2 reduzindo para 0,869. Em contraste, o modelo RF manteve uma consistência entre treino e teste, com R^2 de 0,998 e 0,998 respectivamente, indicando um equilíbrio mais estável e generalização superior. O modelo ANN, por sua vez, mostrou um R^2 de 0,674 no treino e 0,717 no teste, o que reflete um desempenho moderado, sem indícios claros de *overfitting*, mas com espaço para melhorias na acurácia. De forma geral, o modelo RF destaca-se como o mais satisfatório em termos de consistência e generalização.

Tabela 2 Resultados das métricas de avaliação dos modelos para treino e teste

Modelo	Treino			Teste		
	MAE	RMSE	R^2	MAE	RMSE	R^2
KNN	0,000	0,000	1,000	0,395	0,571	0,869
RF	0,027	0,063	0,998	0,040	0,075	0,998
ANN	0,755	0,912	0,674	0,718	0,8399	0,717

A Figura 7 demonstra o desempenho dos modelos KNN, RF e ANN, por meio de gráficos de dispersão comparando valores reais e previstos nos dados de teste. Para o KNN, observa-se uma alta correlação entre as previsões e os valores reais, embora exista espaço para melhorias. O modelo RF alcança um desempenho de R^2 excelente. Por fim, o modelo ANN denota uma capacidade moderada de previsão. As diferenças entre os gráficos de dispersão dos modelos realçam a variabilidade em sua capacidade de modelar a relação entre as variáveis dependente e independente no conjunto de dados em estudo.

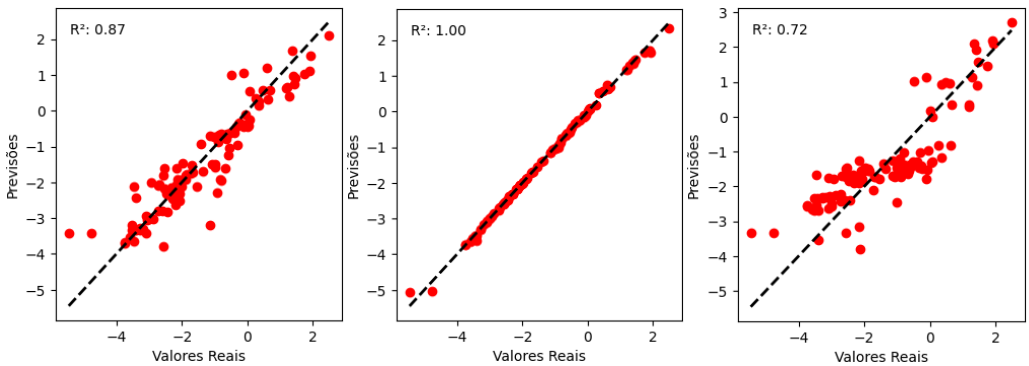

Figura 7 Desempenho dos modelos nos dados de teste, KNN, RF e ANN, respectivamente.

4 CONCLUSÃO

A fim de desenvolver um modelo preditivo eficaz para a estimativa do coeficiente de permeabilidade do solo, aplicou-se técnicas avançadas de aprendizado de máquina, incluindo KNN, RF e ANN. O refinamento do banco de dados foi uma etapa crucial, em que se optou pelo logaritmo do coeficiente de permeabilidade para treino e teste, excluindo-se valores atípicos e registros incompletos, resultando em um conjunto de dados robusto com 703 entradas válidas. A análise estatística revelou correlações variáveis entre os parâmetros, com a maioria exibindo uma relação linear positiva, embora fraca a moderada.

Os modelos foram rigorosamente avaliados por meio de métricas de desempenho. O modelo KNN, apesar de exibir uma adaptação perfeita aos dados de treino, sugeriu potencial *overfitting*, evidenciado por uma redução significativa no desempenho no conjunto de teste. O modelo RF demonstrou uma excelente consistência e capacidade de generalização, com um coeficiente de determinação próximo da excelência nos dados de teste. Por outro lado, o modelo ANN apresentou um desempenho satisfatório, sem sinais significativos de *overfitting*, mas com margem para melhorias em termos de precisão.

Os resultados do estudo indicam que o modelo RF é o mais robusto e confiável, proporcionando um equilíbrio entre acurácia e generalização, tornando-se o mais indicado para a estimativa do coeficiente de

permeabilidade dentro do escopo deste trabalho. As diferenças nos desempenhos dos modelos também sublinham a importância de uma avaliação cuidadosa para a escolha do algoritmo mais adequado para tarefas preditivas específicas em geotecnia.

Cabe ressaltar que existe margem para evolução desses algoritmos com uma generalização maior do banco de dados. Ao incluir materiais mais finos e uma quantidade maior de dados pode-se chegar a indicativos das principais variáveis de influência na permeabilidade e quantificar a contribuição de cada uma dessas variáveis e, assim, contribuir para uma teoria geral de permeabilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

Cedergren, H. R. (1989). *Seepage, Drainage, and Flow Nets*. Wiley.

Feng, S., Barreto, D., Imre, E., Ibraim, E., & Vardanega, P. J. (2023). Use of hydraulic radius to estimate the permeability of coarse-grained materials using a new geodatabase. *Transportation Geotechnics*, *41*, 101026. https://doi.org/10.1016/j.trgeo.2023.101026

Filho, D. B. F., & Júnior, J. A. S. (2009). Desvendando os Mistérios do Coeficiente de Correlação de Pearson. *Revista Política Hoje*, 115–146.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (N. Tache, Ed.; Second Edition). O'Reilly Media, Inc.

Hazen, A. (1892). *Some Physical Properties of Sands and Gravels, with Special Reference to Their Use in Filtration*. 24th Annual Report, Massachusetts State Board of Health, 539-556.

Harr, M. E. (1991). *Groundwater and Seepage*. Dover.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction of Statistical Learning with Applications in Python*.

Kanungo, D. P., Sharma, S., & Pain, A. (2014). *Artificial Neural Network (ANN) and Regression Tree (CART) applications for the indirect estimation of unsaturated soil shear strength parameters*. Frontiers of Earth Science, 8(3), 439–456. https://doi.org/10.1007/s11707-014-0416-0

Lebron, I., Schaap, M. G., & Suarez, D. L. (1999). *Saturated hydraulic conductivity prediction from microscopic pore geometry measurements and neural network analysis*. Water Resources Research, *35*(10), 3149–3158. https://doi.org/10.1029/1999WR900195

Mello, V. F. B., & Teixeira, A. H. (1962). *Mecânica dos Solos*. In Universidade de São Carlos (Vol. 1).

Mitchell, J. K., & Soga, K. (1930). *Fundamentals of Soil Behavior* (INC. JOHN WILEY & SONS, Ed.; Third Edition).

Rawls, W. J., & Brakensiek, D. L. (1989). *Estimation of Soil Water Retention and Hydraulic Properties*. Unsaturated Flow in Hydrologic Modeling, 275–300.

Sperry, J. M., & Peirce, J. J. (1995). *A Model for Estimating the Hydraulic Conductivity of Granular Material Based on Grain Shape, Grain Size, and Porosity*. Groundwater, *33*(6), 892–898.